Multiple layers of contrasted images for robust feature-based visual tracking

Xi Wang, Marc Christie, Eric Marchand

▶ **To cite this version:**

Xi Wang, Marc Christie, Eric Marchand. Multiple layers of contrasted images for robust feature-based visual tracking. ICIP'18 - 25th IEEE International Conference on Image Processing, Oct 2018, Athens, Greece. pp.241-245, 10.1109/ICIP.2018.8451810 . hal-01822789

HAL Id: hal-01822789
https://inria.hal.science/hal-01822789

Submitted on 25 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MULTIPLE LAYERS OF CONTRASTED IMAGES FOR ROBUST FEATURE-BASED VISUAL TRACKING

Xi Wang, Marc Christie, Eric Marchand

Univ Rennes, Inria, CNRS, IRISA

ABSTRACT

Feature-based SLAM (Simultaneous Localization and Mapping) techniques rely on low-level contrast information extracted from images to detect and track keypoints. This process is known to be sensitive to changes in illumination of the environment that can lead to tracking failures. This paper proposes a multi-layered image representation (MLI) that computes and stores different contrast-enhanced versions of an original image. Keypoint detection is performed on each layer, yielding better robustness to light changes. An optimization technique is also proposed to compute the best contrast enhancements to apply in each layer. Results demonstrate the benefits of MLI when using the main keypoint detectors from ORB, SIFT or SURF, and shows significant improvement in SLAM robustness.

***Index Terms*—** keypoint detection, contrast enhancement, SLAM

1. INTRODUCTION

Research in visual tracking systems such as SLAM and SfM (Structure from Motion) has led to mature technologies exploited in industrial-level systems. Except for *direct methods* working on the analysis of changes in pixel gradients, the majority of visual SLAMs rely on corner detection with *extractors* that extract keypoints (KP) and *descriptors* that identify and match the extracted KPs over different frames.

Unfortunately, the corner detection process and consequently the matching problem are strongly dependent on the illumination condition at the moment of capturing images and generally make a brightness constancy assumption. Although the matching process usually relies on gradient information that is more or less independent from intensity, SLAM and SfM methods still suffer from illumination changes at different degrees (see Fig. 1) and may yield inaccurate maps and even tracking failures during the tracking process [1, 2].

Robustness to light changing conditions is therefore a central issue that has received increased attention. The issue has often been tackled at the *extractor* level by searching an optimal contrast threshold in the KP extractor with respect to the current lighting condition. For example, in SuperFast [3] the FAST contrast threshold – a threshold value that triggers a brighter, darker or similar decision on per-pixel comparison – is dynamically computed using a feedback-like optimization method that yields a new threshold value per region in the image. Lowering the threshold however tends to generate a large number of KPs that influence the computational capacity of other processes, and the proposed technique requires specific adaptations to be applied to other KP detectors.

Fig. 1. ORB keypoint extractor and descriptor on different lighting conditions (images a and b). Only a few keypoints are matched between (a) and (b) using ORB (*i.e.* same position, same descriptor), compared to our MLI method (image d).

Another possibility is to apply image transformations (*eg.* contrast enhancers) on captured images before applying KP detectors. Interestingly, it has been demonstrated that KP extractors gain significant performance by using HDR (High Dynamic Range) images as input, converted to SDR (Standard Dynamic Range) images through tone-mapping operators [4, 5]. Among these techniques, a learning-based optimal tone-mapping operator has been proposed for SIFT-like detectors [6]. But the high computational cost and specific HDR devices are required, as well as HDR-customized extractors hamper the wider applicability of such approaches. In comparison, for SDR images, research has mainly focused on contrast enhancement operators for aesthetic and perceptional goals through changes in the exposure times [7] which remain limited in addressing robustness of KP tracking.

For direct and semi-direct SLAM methods, *i.e.* methods that rely on analysis of pixel intensities rather than extracting intermediate features, robustness to illumination changes has been addressed by optimizing an affine brightness transfer transformation between consecutive frames [8, 9]. Using mutual information instead of photometric error as the metric during the optimization process of pose estimation has also demonstrated its benefits [10, 1]. While exhibiting a good robustness to illumination changes, these methods remain computationally expensive.

In this paper, a new representation – multi-layered image (MLI) – is proposed to improve the robustness of visual tracking systems like SLAM. We first analyze the reasons for which KP extraction fails. We then present our MLI representation that relies on different contrast-enhanced versions of the same image. We design an optimization technique to compute the best contrast enhancements to perform and we demonstrate how MLI boosts performance of traditional KPs extractors/descriptors, and lead to significant improvements on the robustness of a state-of-the-art visual SLAM system.

2. ISSUES WITH CONTRAST ENHANCERS

Except for HDR images, the majority of contrast enhancement techniques can be defined as continuous monotone surjective mappings from domain interval $[0, 1]$ to codomain interval $[0, 1]$ that transform a given image to a (more) contrasted version. Typically, the classical S-Curve Tone Mapping method [7] is used to correct underexposure and overexposure regions in images, by applying a per-pixel function.

However, by using such transformations the improvement of the contrast in one region must necessarily be *paid* for by a reduction of the contrast in another region (see examples of S-Curves in Fig. 2). Given that most keypoint detection techniques are based on analysis of finite local differences in contrasts, contrast enhancement tends to increase the detection of keypoints by passing some internal thresholds, while contrast compression leads to the opposite. We illustrate this through the example of a synthetic scene shot in different lighting conditions [11]. The ORB detector [12] is used to extract and match keypoints between a well-lit given reference image (see Fig. 1 a) and an image from the same viewpoint with only a flashlight illuminating the scene. The ORB detector with default parameters only finds a few *matched keypoints* between the two images (*i.e.* keypoints extracted and described as being the same at the same image locations). Interestingly by applying different S-Curve transformations (see Fig. 2), the total number of keypoint matches increases while ORB detector loses already matched keypoints from previous transformations. This empirically shows (i) that a single contrast enhancer only represents a partial solution to keypoint robustness, and that (ii) improving extraction by contrast-enhancement also improves matching.

3. MULTI-LAYERED IMAGE

The idea of Multi-Layered Image (MLI) is to generate k contrast enhancements of a given image into k image layers on which keypoint detection will be performed. The contrast enhancement technique relies on a saturated affine brightness transfer per-pixel function (SAT). We use a SAT form that defines a *contrast band* $\mathbf{u} = (a, b)^\top$ which conveniently models the lower cut point (a) and higher cut point (b) of the saturation, with a linear interpolation between a and b on pixel

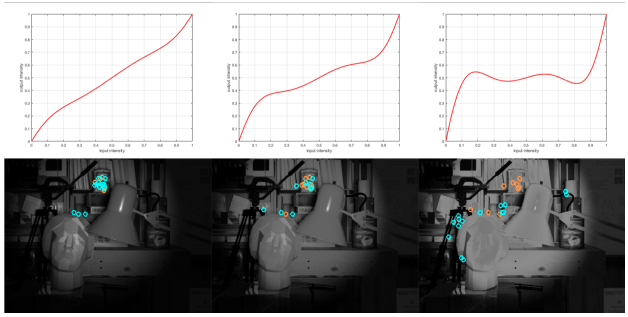

Fig. 2. Matching KPs with ORB detector between a reference image (see Fig. 1 a) and different S-Curve tone mapped versions of an image in a different lighting condition. Each tone-mapping provides newly matched KPs (blue) while losing others (orange).

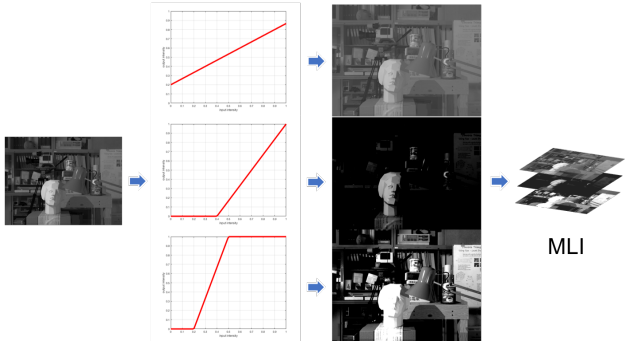

Fig. 3. Using SAT function with different contrast bands to generate a multi-layered image representation (MLI).

intensity i (see Fig. 3). A given contrast $\mathbf{u} = (a, b)^\top$ is defined in a contrast space $\Gamma \subseteq \mathbb{R}^2$, where Γ is the space of all contrast bands where $b > a$.

$$f_{SAT}(i, (a, b)^\top) = min(max(0, i/(b - a)), 1) \qquad (1)$$

Parameters a and b naturally represent the *band* region where the contrast is enhanced, which motivated the choice of this operator compared to S-Curve. To ensure enhancement or compression of contrasts, we define the range of values for $\mathbf{u} = (a, b)^\top$ as $a \in [-\infty, 1]$ and $b \in [0, \infty]$. The computation of a layer k in our MLI representation is performed by applying the following operator MLI_k on all pixel intensities of the image using a contrast band $\mathbf{u}_k$. A MLI is therefore represented as a set of k image layers where $MLI_k(I) = f_{SAT}(I, \mathbf{u}_k)$ for an image I, where $f_{SAT}(I, \mathbf{u}_k)$ is the application of f_{SAT} on all pixels of I.

4. LOW-CORRELATED CONTRAST SPACE

To address the issue of robustness, the key challenge is therefore to generate different layers such that each layer has the *lowest correlation* with the others in terms of detected keypoints (*i.e.* aiming at providing new keypoints in each layer).

In other terms, we are looking at computing a set of contrast band parameters such that each contrast band yields an image containing newly matched keypoints with the reference image (the initial lighting condition).

This paper proposes a technique to compute the optimal contrast bands together with a stopping criterion on the number of layers required, given a reference image I^* representing a given lighting condition and a camera image I in another lighting condition. The first layer is computed by selecting a contrast band $\mathbf{u}_i$ that maximizes the *correspondence* of keypoints between the reference image I^* and the contrast-enhanced image $f_{SAT}(I, \mathbf{u}_i)$. The other layers are computed by selecting contrast bands that provide the lowest correlation (in terms of correspondence between keypoints) with the current contrast band. More formally we start by defining the *KP-correspondence* set between two images. Given S^* the set of KPs extracted from a reference image I^* (and respectively S from I), the *KP-correspondence* S_{Cor} is the set of KPs in S^* for which there is a *correspondence* in S, *i.e.* for which there is a keypoint at a similar location in image I:

$$S_{Cor} = \{x^* \,|\, x^* \in S^*, x \in S, \|x^* - x\| < \epsilon\} \tag{2}$$

This definition expresses the *repeatability* ratio [13] between two sets of keypoints from two different images, a well-known metric in visual tracking [14], that measures the ratio of KPs appearing at similar positions on both images, over the number of keypoints in the first image:

$$\frac{Card(\{x^* \in S^*, x \in S,\ s.t.\ \|x^* - x\| < \epsilon\})}{Card(S^*)} \tag{3}$$

Given a KP extractor e, we define a *band-correspondence set* $S^{\mathbf{u}}_{Cor}$ as the *KP-correspondence* set between a SAT contrast-enhanced version of I and a reference image I^* given $\mathbf{u}$:

$$S^{\mathbf{u}}_{Cor} = \{x^* \,|\, x^* \in e(I^*), x \in e(f_{SAT}(I, \mathbf{u})), \|x^* - x\| < \epsilon\} \tag{4}$$

Intuitively, this means the more keypoints there are in this *band correspondence set*, the better is the contrast band $\mathbf{u}$ in yielding an image containing corresponding keypoints with a reference image. We therefore express the cardinality of this set $M_{Cor} : \Gamma \subseteq \mathbb{R}^2 \rightarrow \mathbb{R}$ as $M_{Cor}(\mathbf{u}) = Card(S^{\mathbf{u}}_{Cor})$. The global maximum of this function represents the optimal contrast band $\mathbf{u}$ in terms of *KP-correspondence* and is used to compute the first layer of our MLI.

We then need a way to compute new contrast bands with low-correlation in the contrast space $\mathbf{u} \in \Gamma$. We define a covariance-like method on $S^{\mathbf{u}}_{Cor}$ that provides a *co-Correspondence* set $S^{\mathbf{u}_1,\mathbf{u}_2}_{coCor}$. This *co-Correspondence* computes the corresponding keypoints between two contrast bands $\mathbf{u}_1$ and $\mathbf{u}_2$ and a reference image I^*. We similarly define its cardinality $M_{coCor} : \Gamma \times \Gamma \subseteq \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$.

$$S^{\mathbf{u}_1,\mathbf{u}_2}_{coCor} = \{x_1 \in S^{\mathbf{u}_1}_{Cor} \,|x_2 \in S^{\mathbf{u}_2}_{Cor}, \|x_1 - x_2\| < \epsilon\} \tag{5}$$

Algorithm 1 Optimal MLI

$i \leftarrow 0$; $C^0(\mathbf{u}) \leftarrow M_{Cor}(\mathbf{u})$;
while $i = 0$ **or** $C^i(\mathbf{u}_i) > k * C^{i-1}(\mathbf{u}_{i-1})$ **do**
  $\mathbf{u}_i \leftarrow \text{argmax}_{\mathbf{u}}(C^i(\mathbf{u}))$
  $C^{i+1}(\mathbf{u}) \leftarrow C^i(\mathbf{u}) - M^{\mathbf{u}_i}_{sim}(\mathbf{u})$
  $i \leftarrow i + 1$
end while
return $\{\mathbf{u}_k\}_{k=1..N}$

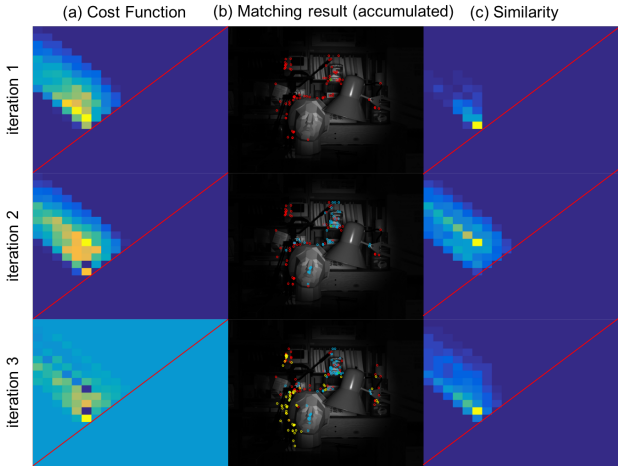

Fig. 4. Evolution of MLI layers: (a)(c) represent heatmaps of cost function $C^i(\mathbf{u})$ and similarity $M^{\mathbf{u}_i}_{sim}(\mathbf{u})$ in each iteration. (b) demonstrates *accumulated* matched ORB KPs against reference image from every iteration represented with different colors. In each heatmap, vertical and horizontal axis represents $\mathbf{u} = (a, b)^\top$ respectively with $a < b$.

$$M_{coCor}(\mathbf{u}_1, \mathbf{u}_2) = Card(S^{\mathbf{u}_1,\mathbf{u}_2}_{coCor}) \tag{6}$$

Using this co-correspondence definition we can compute how much a given contrast-band $\mathbf{u}_r$ yields information similar to all the other contrast bands, *i.e.* the number of KPs generated by this contrast-band also found in others. This similarity measure $M^{\mathbf{u}_r}_{sim} : \Gamma \subseteq \mathbb{R}^2 \rightarrow \mathbb{R}$ is expressed as:

$$M^{\mathbf{u}_r}_{sim}(\mathbf{u}) = Card(S^{\mathbf{u}_r,\mathbf{u}}_{coCor}) \tag{7}$$

A low similarity represents a low correlation between the contrast bands. Using these definitions, the computation of the different contrast bands consists in applying a sequence of two stage operations (see Alg. 1). The first stage selects an optimal contrast band $\mathbf{u}$ that maximizes a cost function $C(\mathbf{u})$, *i.e.* maximizes the correspondences between reference image I^* and $f_{SAT}(I, \mathbf{u})$. The second stage then updates the cost function by subtracting the similarity $M^{\mathbf{u}_i}_{sim}$ between the current contrast band $\mathbf{u}_i$ and all others (ensuring a low correlation). The algorithm terminates when the new iteration yields information proportionally lower than the previous one using a factor k.

cam \ ref		Daylight			Fluorescent			Lamps			Flashlight		
		ORB	SIFT	SURF	ORB	SIFT	SURF	ORB	SIFT	SURF	ORB	SIFT	SURF
Daylight	D	100/100	100/100	100/100	63.6/21.2	36.2/21.5	50.1/20.0	21.1/0.8	22.2/0.5	28.6/1.1	52.3/5.8	43.0/11.3	48.6/5.9
	M	100/100	100/100	100/100	**85.3/38.7**	**64.1/37.3**	**75.4/34.4**	**35.7/1.6**	**39.8/1.1**	**49.8/2.2**	**67.6/8.4**	**50.9/13.8**	**56.7/8.4**
Fluorescent	D	63.7/21.2	48.7/27.2	63.2/22.8	100/100	100/100	100/100	7.0/0.3	33.3/1.0	44.4/1.5	49.8/9.1	54.5/16.9	59.7/8.4
	M	**72.3/34.7**	**61.2/34.7**	**76.6/30.7**	100/100	100/100	100/100	**13.8/0.4**	**51.0/1.6**	**65.7/2.0**	**66.2/13.9**	**60.8/21.5**	**65.4/13.4**
Lamps	D	4.2/0.9	2.0/1.2	3.1/1.7	1.4/0.5	2.1/1.4	3.6/1.7	100/100	100/100	100/100	4.4/1.0	1.2/0.5	3.8/0.8
	M	**64.6/20.0**	**46.9/24.0**	**62.1/21.8**	**66.6/21.8**	**47.0/26.7**	**60.7/25.9**	100/100	100/100	100/100	**56.8/6.7**	**41.3/14.2**	**46.2/9.0**
Flashlight	D	34.0/5.3	12.2/5.4	16.4/5.3	32.4/7.6	11.3/6.3	16.2/6.2	14.6/0.5	5.1/0.0	12.1/0.2	100/100	100/100	100/100
	M	**58.1/11.8**	**31.4/11.8**	**44.6/11.4**	**61.4/16.6**	**30.5/15.0**	**44.0/13.5**	**18.4**/0.5	**16.4/0.5**	**35.0/0.8**	100/100	100/100	100/100

Table 1. Repeatability/matching ratio evaluation between MLI (M) and default single image (D) in percentage.

The algorithm is illustrated in Fig. 4 using the FAST extractor [15], and the ground truth is ensured by calculating BRIEF descriptor [16]. Again we reuse the data set of New Tsukuba [11]. For the purpose of illustration, Γ is defined as a discrete sampling space between $[-0.5, 1.5] \times [-0.5, 1.5]$. Images (a) and (c) represent the landscape of the cost functions $C^i(\mathbf{u})$ and similarities $M_{sim}^{\mathbf{u}_i}(\mathbf{u})$ of the previous optimal contrast band in each iteration. We observe that the maximums of $C^i(\mathbf{u})$ change every iteration after updating by a subtraction with M_{sim}. As one can see in the third iteration (Fig. 4.b), extract more KPs with low-correlation between the layers. This empirically shows that instead of intuitively or programmatically decreasing the thresholds parameters of detectors, an optimization scheme to compute the optimal contrast bands in each layer improves the correspondence of keypoints with a reference image.

5. EVALUATIONS AND EXPERIMENTS

We compare the use of MLI with classical detectors/descriptors: ORB [12], SIFT [17] and SURF [18] on the NewTsukuba Data set [11]. We measure the *repeatability* (*cf.* eq. 3) and the *matching* ratio by matching descriptors on feature points between a reference image I^* and new images I from the same viewpoint and different lighting conditions. For each condition, the optimal values of the contrast bands are computed by using the algorithm defined in Alg 1. The measures reported in Table 1 show that MLI improves repeatability and matching ratio across all detectors, despite different detection methods and default threshold parameters. This demonstrates the wide applicability of our approach.

We then compare the use of MLI on visual SLAM tasks in different lighting conditions. We choose ORB-SLAM [19] to implement our MLI representation. We tested two sequential videos from a combination of four different lighting conditions. The experiment consisted in localizing and tracking the camera from the second video sequence against the keyframes generated from the first video sequence (in a way similar to NID-SLAM [1]). The measured value is the success rate, *i.e.* the percentage of the frames from second video successfully tracked against keyframes created from the first video.

The optimal contrast bands of the MLIs of each illumination condition (first video to second video) are computed by Algo. 1 over 5 sample images in the test set. Comparison is performed between standard ORB-SLAM [19], our MLI implemented ORB-SLAM and reported results of monocular visual SLAM NID-SLAM [1] which demonstrated a good performance against illumination changing environments. Two to three layers are used in the experiments. The MLI approach is significantly more robust than the default ORB implementation (see Table 2), especially in difficult situations (Lamps to Flashlights, Daylight to Flashlight). Our approach also compares very favourably with NID-SLAM, displaying similar or better performances for a lower computational cost (keypoint extraction only represents 3 to 5% of computation time in ORB [19], limiting the impact of MLI cost). Typically the Lamps to Flashlight failed to track all keyframes with both ORB and NID (0%) and successfully tracked 94.2% of the keyframes with our MLI approach.

V_2 \ V_1	Daylight			Fluo			Lamps			Flash		
	NID	ORB	MLI	NID	ORB	MLI	NID	ORB	MLI	NID	ORB	MLI
Daylight	99.3	100	**100**	96.7	96.2	**98.4**	73.9	**97.6**	53.6	74.6	**79.8**	77.1
Fluo	95.0	88.1	**95.1**	99.7	100	**100**	85.3	93.9	**100**	95.8	100	**100**
Lamps	88.3	55.7	**93.3**	**93.6**	79.8	93.4	93.1	100	**100**	84.3	37.9	**96.8**
Flash	23.8	30.7	**77.6**	92.2	90.6	**93.6**	0.00	0.00	**94.2**	92.0	**100**	99.3

Table 2. SLAM keyframe retrieval success rate between our MLI implementation, default ORB SLAM and NID SLAM.

6. CONCLUSION

We have introduced a novel multi-layered image representation to address robustness of keypoint tracking in light changing conditions. Each layer is a contrast-enhanced version of an original image, computed in a way to improve the detection and matching of keypoints. We proposed a method to compute optimal parameters for each layer and demonstrated the benefits of our approach on keypoint detection tasks and SLAM applications. While the proposed solution pre-computes the optimal contrast bands to apply, we are currently exploring an adaptive computation of such transforms.

7. REFERENCES

[1] G. Pascoe, W. Madder, M. Tanner, P. Piniés, and P. Newman, "Nid-slam: Robust monocular slam using normalised information distance," in *Conf. on Computer Vision and Pattern Recognition*, 2017.

[2] P. Seonwook, S. Thomas, and P. Marc, "Illumination change robustness in direct visual slam," in *2017 IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2017.

[3] G. Florentz and E. Aldea, "Superfast: Model-based adaptive corner detection for scalable robotic vision," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2014, pp. 1003–1010.

[4] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in hdr based imaging under changes in illumination conditions," in *IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 289–294.

[5] B. Přibyl, A. Chalmers, and P. Zemčík, "Feature point detection under extreme lighting conditions," in *Proceedings Spring Conf. on Computer Graphics*. ACM, 2013, pp. 143–150.

[6] A. Rana, G. Valenzise, and F. Dufaux, "Learning-Based Tone Mapping Operator for Image Matching," in *IEEE Int. Conf. on Image Processing (ICIP)*, Beijing, China, Sept. 2017, .

[7] J. Yuan, L. Sun, "Automatic exposure correction of consumer photographs," *European Conf. on Computer Vision (ECCV)*, pp. 771–785, 2012.

[8] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[9] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*, September 2014.

[10] G. Caron, A. Dame, and E. Marchand, "Direct model based visual tracking and pose estimation using mutual information," *Image and Vision Computing*, vol. 32, no. 1, pp. 54–63, 2014.

[11] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui, "Towards a simulation driven stereo vision system," in *Proceedings of Int. Conf. on Pattern Recognition (ICPR)*, Nov 2012, pp. 1038–1042.

[12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[13] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Int. J. of computer vision*, vol. 37, no. 2, pp. 151–172, 2000.

[14] S. Gauglitz, T. Höllerer, and M. Turk, "Evaluation of interest point detectors and feature descriptors for visual tracking," *Int. J. of Computer Vision*, vol. 94, no. 3, pp. 335, Mar 2011.

[15] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *European Conf. on Computer Vision (ECCV)*, pp. 430–443, 2006.

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *European Conf. on Computer Vision (ECCV)*, pp. 778–792, 2010.

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conf. on Computer Vision (ECCV)*, pp. 404–417, 2006.

[19] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Trans. on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.